Running head: INDIVIDUAL DIFFERENCES IN INTERPERSONAL ACCURACY

Individual Differences in Interpersonal Accuracy:

A Multi-level Meta-Analysis to Assess Whether Judging Other People is One Skill or Many

Katja Schlegel

Northeastern University

R. Thomas Boone

University of Massachusetts, Dartmouth

Judith A. Hall

Northeastern University

Author Note

Correspondence concerning this article should be addressed to R. Thomas Boone, Department of Psychology, University of Massachusetts, Dartmouth, 285 Old Westport Road, North Dartmouth, MA, 02747, USA. Email: tboone@umassd.edu .

Acknowledgements

The authors thank Curtis Pitegoff for bibliographic searching and data entry, and the many authors who generously provided their unpublished data.

Abstract

Research into individual differences in interpersonal accuracy (IPA; the ability to accurately judge others' emotions, intentions, traits, truthfulness, and other social characteristics) has a long tradition and represents a growing area of interest in psychology. Measuring IPA has proved fruitful for uncovering correlates of this skill. However, despite this tradition and a considerable volume of research, very few efforts have been made to look collectively at the nature of the tests involved in assessing IPA, leaving questions of the broader structure of IPA unresolved. Is IPA a single skill or a clustering of many discrete skills or some combination of partially overlapping skills? In a multi-level meta-analysis of 103 published and unpublished participant samples (13,683 participants), we analyzed 622 correlations between pairs of IPA tests (135 different IPA tests altogether). The overall correlation between IPA tests was $r = .19$, corrected for the nesting of correlations within the studies that administered more than two IPA tests and reported several correlations for the same participant sample. Test domain and characteristics were evaluated to explain differences in effect sizes; in general, tests in similar domains and using similar methodologies were more highly correlated with each other, suggesting that there are domains within which individual differences cluster. Implications for future research and IPA measurement were discussed.

Keywords: individual differences, interpersonal accuracy, emotion recognition, lie detection, personality judgment, meta-analysis

Individual Differences in Interpersonal Accuracy:
A Multi-level Meta-Analysis to Assess Whether Judging Other People is One Skill or Many

The ability to make correct inferences about other people's states and traits is a crucial interpersonal skill. In the present article, we use the term *interpersonal accuracy* (IPA) to encompass the entire breadth of ways in which people can be accurate in perceiving others' characteristics based on exposure to their behavior or appearance. IPA is meant to constitute a superordinate ability that subsumes specific judgment skills related to emotion, deception, personality, and many other transient and enduring characteristics of people. Over the years, researchers who are interested in accurate interpersonal perception have used a plethora of terms including *interpersonal sensitivity* (Hall & Bernieri, 2001), which is broad enough to encompass considerate social behavior as well as perception accuracy, as well as more specific terms such as *emotion recognition*, which refers to judging discrete emotions (Matsumoto et al., 2000); *empathic accuracy*, which refers to judging another person's thoughts and feelings during spontaneous interaction (Ickes, 2001); *judgmental accuracy*, which refers generally to personality judgment (Funder & Sneed, 1993); and *mental states attribution*, which refers to making judgments about others' intentions and beliefs (Brüne & Schaub, 2012), among many others. *Interpersonal perception accuracy*, referencing accurate perception of any and all states and traits, has been used as a suitable umbrella to subsume many specific terms (Davis & Kraus, 1997; Hall, Schmid Mast, & West, 2016).

IPA encompasses a variety of judgment types that differ on dimensions such as the content domain, spontaneity of encoded behavior, and the sensory cue channel (Hall, Bernieri, & Carney, 2005). Content domains include others' states (such as emotions or deception), traits (such as personality, intelligence, or socioeconomic status), and social characteristics (such as

religion, sexual orientation, or kinship). Spontaneity of the behavior that is being judged can range from spontaneous (e.g., unrehearsed behavior occurring during a social interaction) to deliberate and posed (e.g., prototypical facial expressions posed by actors). A target's behavior and appearance can be presented to the judge (perceiver, decoder) through different cue channels or modalities, such as pictures or videos of faces, body postures, gestures, and vocal recordings. IPA research can be divided into six broad domains, namely Personality (judging others' traits), Emotion (judging others' temporary affective states), Situational Affect (inferring what kind of situation a person is in), Deception (distinguishing truth from falsehood), Thoughts and Feelings (inferring others' spontaneous thoughts and feelings), and Social Attributes (inferring others' social group membership and social characteristics, e.g., kinship).

Researchers have measured IPA for a century (e.g., Adams, 1927; Feleky, 1914), looking at it from many perspectives. The present article focuses on the individual differences perspective—how people differ in their degree of IPA – and aims to analyze how the various types of accuracy are related to each other. IPA research is experiencing a dramatic upward trend, as evident in the trajectory of PsycINFO entries for the term *emotion recognition* over the last few decades: from 87 in the 1990s, to 680 in the 2000s, to over 2,000 in just the half-decade since then. Yet, there has been little progress in understanding the construct of IPA as a whole. One reason is that, understandably, researchers in a given discipline have tended to study the kind of IPA most relevant to them: personality psychologists have studied accuracy of personality judgment (Borkenau & Liebler, 1992; Funder & Colvin, 1988), and social psychologists have tended to study more socially defined accuracies such as in judging sexual orientation or religion (Ambady, Hallahan, & Conner, 1999; Tskhay & Rule, 2013). Similarly, emotion researchers have studied accuracy of judging emotions (Ekman et al., 1987; Gendron,

Roberson, van der Vyver, & Barrett, 2014), as have many researchers in various branches of abnormal psychology who study emotion recognition deficits as a manifestation of emotional or developmental dysfunction (Sayla, Vella, Armstrong, Penn, & Twamley, 2013; Uljarevic & Hamilton, 2013). Although these research programs have generated many valuable results, they have remained largely isolated within their own areas of specialization, meaning that not much effort to understand the broader landscape of individual differences in IPA has occurred.

The present research aimed to produce a more unified picture of IPA by exploring its structure as it is captured collectively in the many measuring instruments that have been used. We accomplished this goal through a multi-level meta-analysis of studies in which researchers administered more than one test of IPA to a group of participants and then correlated the tests together. We did not consider the vast number of studies that used one single IPA test, as these do not allow examining our question of how IPA tests are correlated. Through examining the pattern of correlations among tests in conjunction with information on their characteristics (e.g., their content domain and cue channels), the present meta-analysis asked whether, at the extremes, IPA is one general skill, meaning different tests of all types would have strong convergent validity, or whether there are many independent skills—as many as there are tests to measure such skill; or whether there is a middle ground in which some kinds of tests are more highly correlated with each other than others. In this case the pattern might reveal a coherent structure according to sensible dimensions. The latter possibility could show (for illustration) that skill in judging emotion might be unrelated to skill in judging personality, but different tests of judging emotion (or personality) might measure the same underlying skill.

There is good reason to develop a better understanding of IPA as an individual difference variable that transcends particular research traditions. IPA is important in all social interactions,

from the first encounter with someone to managing a relationship with a close friend, romantic partner, colleague, patient, or child. Higher IPA can make social relationships and interactions more manageable and predictable, and is adaptive from an evolutionary perspective (Ambady & Skowronski, 2008). A wealth of research as well as several meta-analyses point to the adaptive value of IPA in workplaces, clinical settings, social life, and psychological adjustment (Hall, Andrzejewski, & Yopchick, 2009; Elfenbein, Foo, White, Tan, & Aik, 2007; Hall et al., 2016). Some of these correlates suggest causally antecedent experiences such as family environment (Halberstadt, 1986; Hodgins & Koestner, 1993) or training interventions (Blanch-Hartigan, Andrzejewski, & Hill, 2012), and others point to possible causal consequences of IPA, such as being a more effective clinician (Ruben, 2016), learning more in an interpersonal instruction task (Bernieri, 1991), or being a better music teacher (Kurkul, 2007). IPA also informs models of interactions such as those involved in cooperation in social dilemmas and exchange theory, by allowing an individual to gauge whether or not to invest with or against others (Boone & Buck, 2003). Finally, IPA has been included in theoretical models of social intelligence (Thorndike, 1920), interpersonal intelligence (Gardner, 1983), and more recently, emotional intelligence and personal intelligence (Mayer, Panter, & Caruso, 2012; Mayer & Salovey, 1997). In these models, IPA is considered a basic component that allows having more successful, smooth, and adaptive interactions with others.

Yet, little is known about how different types of IPA are related to each other. In describing the many different approaches to measuring IPA, Zebrowitz (2001) used the metaphor of seven blind men, each trying to understand the entirety of an elephant by examining the specific part in front of them, leading each man to infer different qualities about that singular elephant. This lack of coordination has left the broader concept of interpersonal accuracy—the

elephant—as a construct with a yet to be defined structure. Authors have long expressed curiosity, and sometimes concern, over how IPA tests correlate with each other (Cline & Richards, 1960; Crow & Hammond, 1957). Buck (1984) reviewed studies that reported correlations among different IPA tasks and noted generally low correlations. For example, the correlation between the Communication of Affect Receiving Test (CARAT; Buck, 1976) and the Profile of Nonverbal Sensitivity (PONS; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979) was $r = .04$. Hall (2001) reported similar findings, with the correlations between the PONS and the Interpersonal Perception Task (IPT; Costanzo & Archer, 1989) being $r = .20$ and $r = .22$ in two samples, the correlation between the PONS and the CARAT being $r = .16$, and the correlation between the CARAT and the IPT being $r = .10$. These results present an odd paradox: Although these tests all measure accuracy in judging other people and had demonstrated predictive validity with social outcomes and other correlates, they seemed to be nearly independent from each other. What, then, is being measured with such instruments? In the present meta-analysis, several hypotheses regarding the structure of IPA are examined, as mentioned above. In the following, we will discuss some reasons that might support each view.

IPA as a Single Ability

There are several reasons to assume that IPA can be considered a single global ability. First, all IPA measures require paying attention to and processing cues that are emitted by another person through physical appearance and/or verbal and nonverbal behavior. The Lens Model (Brunswik, 1956) suggests that the mechanism through which the state or trait of a sender is encoded and decoded is based on these cues. Across the different IPA domains, there has been a systematic effort to isolate the relevant cues in different media that signal the sender's state or trait of interest in different nonverbal channels including the face, voice, and body (DeMeijer,

1989; Ekman et al., 1987; Izard, 1994; Rule, Ambady, Adams, & Macrae, 2008; Scherer, 1978). Individuals with higher general IPA might be more attentive to all of these cues, developing greater sensitivity to variation and nuance. Other researchers have proposed that there exists a superordinate set of cues relating to spatiotemporal forms, geometric and temporal patterns that underlie cue features in all the nonverbal media (Scherer & Oshinsky, 1977). Recent research has shown that sensitivity to these spatiotemporal cues is related to emotion recognition ability across several media (Castro & Boone, 2015). Although not yet extended to other IPA domains, such sensitivity to spatiotemporal cues may allow the receiver better utilization of a wide range of nonverbal cues and offer a global, or at least shared, underlying mechanism.

IPA as a Collection of Unrelated and Instrument-Specific Skills

This perspective proposes that each IPA test measures a different skill, and that these skills are independent from each other. This view is supported by studies that found very low correlations between different IPA tests. It assumes that the main reason is that IPA measures differ a lot in their content and tasks, even within one content domain, and that these contents and tasks each require independent skills to make an accurate judgment. For example, emotion recognition tests use a wide range of stimuli in which someone expresses an emotional state, including pictures of facial expressions, voice recordings, images of body postures, video clips with or without sound, or combinations of these—and not all types of cues are embedded equally in these stimuli. In addition, diagnostic cues for different emotions tend to appear selectively in specific channels (App, McIntosh, Reed, & Hertenstein, 2011) rather than globally across channels. Thus, a person who is good at judging one type of cues (i.e., static facial expressions) but not others (i.e., dynamic vocal expressions) might perform well only in some emotion recognition tasks and not in others.

Other IPA domains have used a variety of tasks and stimuli as well. Personality judgments are based on still pictures of faces with neutral expressions, voice recordings, or video clips of target persons doing a variety of things (such as introducing themselves, negotiating, or being interviewed for a job), often with the linguistic information maintained in the clips. Deception detection tasks have focused on least two different components involved in catching the liar: an affective component such as anxiety or guilt, or a cognitive component where the would-be liar must manage the complexity of telling a lie that lines up with all the details of the truth (DePaulo, LeMay, & Epstein, 1991). Arguably, different IPA tests represent a combination of tasks relying on different types of information and might therefore involve quite distinct processes. Some tests, such as the IPT (Costanzo & Archer, 1989), were designed to have heterogeneous content (e.g., judgments on kinship, competition, or status), in an effort to span a wider range of social situations. Heterogeneous content and different cue channels also require knowledge about which cues to pay attention to for various types of judgments. Individuals might specialize only in some types of content or judgments, leading to lower associations between measures that differ in these respects.

Even within one cue channel, such as the face, different mechanisms might be involved in inferring information from someone's facial features (e.g., thick brows, plump lips) versus his or her emotional expressions, which involve muscle movements (Ekman & Friesen, 1978). Furthermore, within one cue channel, stimuli differ in the spontaneity of encoded behavior and in stimulus dynamism. Spontaneity can range from spontaneous (e.g., unrehearsed behavior recorded during a social interaction) to deliberate and posed (e.g., facial expressions posed by actors).

Each way in which stimuli in IPA tasks can differ might require distinct skills from the judge that draw on different types of cognitive processing, such as more superficial and automatic processes versus more conscious and resource-consuming processes. These skills might be quite independent, thus leading to no specific pattern between the different IPA measures even within a content domain. Nevertheless, each of the various tasks and ways to measure IPA might still represent a justifiable and valid assessment approach that captures a different facet of the broader concept of IPA. Thus, accuracy is comprised of modular units based upon discrete skills in an additive process rather than overlapped skills with a shared process.

IPA as a Set of Distinct but Correlated Skills

This perspective assumes that certain types of IPA tests are substantially correlated with each other, but that different types of tests are related to a lower extent. Such types or clusters of IPA tests might be defined by similar content (e.g., judging deception), similar cue modalities, or similar types of cue processing more generally. Each type might require similar cognitive decision-making processes, whereas different types of tests or methodologies might have less in common. From this perspective, IPA would be a multi-faceted, hierarchical construct in which all IPA clusters or facets share some variance due to a global IPA skill, but in which connections between clusters vary according to how many features the respective tests have in common. Several distinctions on the level of features and cue processing can be made between IPA tasks. First, interpersonal judgments can refer to others' transient states (such as emotions or deception) and others' more enduring qualities (such as personality, intelligence, social status, or age). In addition, judgments can be about continuous evaluations (How extraverted is the person?), categorical choices (Which out of five emotions did the person express?), or dichotomous

decisions (Did this person tell the truth or not?). Interpersonal judgments also differ in how the stimuli are presented, (e.g., static photographs vs. dynamic video, posed vs. spontaneous expressions). Each of these methodologies might represent a type of IPA within which tests are more highly correlated.

In addition, the different types of judgments might also be influenced by different individual characteristics, traits, and self-perceptions of the judge. For example, it would be plausible to assume that a strong belief in justice or certain occupational requirements might motivate someone to be more focused on being an accurate judge of deception and more likely to acquire that skill over time, but not as motivated in judging social attributes. Similarly, it might be that people who rate themselves as high on emotional intelligence would be better at recognizing others' emotions, but not at evaluating others' extraversion. Or, people who work on a suicide hotline might become very skilled in reading vocal cues, but have only average skill in reading cues in other channels. Indirect support for this model of IPA comes from research on how other variables are related to IPA, across different ways of measuring it. Gender differences have been particularly well researched in this regard, with several meta-analyses available. If IPA were one global, homogeneous skill one might expect gender differences to be consistent regardless of how IPA was defined in a given test. However, gender differences are inconsistent, with women having more accuracy than men for judging emotions (e.g., Thompson & Voyer, 2014) and for judging personality (Hall, Gunnery, & Horgan, 2016; Vogt & Colvin, 2003), but no gender difference is evident for distinguishing between truth-telling and lying (Aamodt & Custer, 2006).

The Present Meta-Analysis

To summarize, there are a number of arguments for each of the three possible perspectives on the structure of IPA. The present meta-analysis aimed to determine which perspective has the most empirical support by examining the magnitude and pattern of correlations between different IPA tests. It includes all the studies we could locate, published and unpublished, in which two or more different tests of IPA were administered and correlated with each other. In addition to establishing the overall correlation between all IPA tests, we also investigated whether the magnitude of these correlations varied depending on characteristics of the IPA tests as well as on features of the specific study and sample. For each IPA test, the content domain it covered was coded, as well as the cue channels, whether stimuli were presented in a static or dynamic mode, whether the stimuli had been created from posed or spontaneous expressions, which response format it used, and whether the test was a standard IPA measure or not. These variables allowed investigating whether there are types of IPA tests and methodologies within which people's levels of accuracy are more highly correlated. Other variables that were examined as potential moderators of the association between IPA tests include the internal consistency of each test, year and publication status of the original study, and sample characteristics (e.g., gender composition).

Method

Search

To find studies that administered two or more IPA tests, several search methods were used. PsycINFO was searched for names of authors known to conduct IPA research, as well as for names of commonly used tests. The terms interpersonal accuracy, nonverbal sensitivity, nonverbal decoding, emotion recognition, personality judgment, and other similar terms were also searched. For both published studies and dissertations, the abstracts, and also full text when necessary, were read to see if two or more tests were used. If, in any study, two or more tests

were used but there was no report of the correlation between them, the study's authors were sent a request for the correlation(s). Also, authors were queried if significant methodological information was missing. The bibliographies of retrieved works were also examined, and the present authors' own reprint files and data archives were consulted. Individual authors known by the present authors to conduct IPA research were emailed with a request for published or unpublished studies, and an announcement was sent to the listservs of the Society for Personality and Social Psychology and the International Society for Research on Emotions.

Inclusion Criteria

Throughout, the term "participants" refers to the sample of test-takers in the studies retrieved for the meta-analysis. To be included, the following criteria needed to be met:

(1) Reported in English.

(2) Participants at least 14 years old on average.

(3) Participant sample size at least 10.

(4) Two or more tests of IPA were given. The definition of an IPA test was that participants viewed and/or listened to recordings or photographs of people and made judgments of some state, trait, or personal attribute, and their judgments were scored for accuracy. Scoring could be for total accuracy (i.e., all the items on a given test) and/or for subparts (see next section for further information on subparts). Studies were excluded if they were based on live dyadic interaction in which, for example, one member of a dyad made judgments of a partner's states or traits and these were scored for accuracy against criteria supplied by the partner (e.g., emotions). This exclusion was done because studies with this design are typically confounded to an unknown degree by unmeasured variation in the partner's accuracy of expressing their state or trait (cf. Hall, Schmid Mast, & Latu, 2015), a problem not present when groups of participants make judgments of a common set of target persons (stimuli).

(5) The original author correlated the tests with each other (or we were able to obtain such correlations; see above).

As stated above, for any given sample of participants, the author may have reported correlations for a total score and/or for subparts (i.e., subsets of items). All of these were allowed for inclusion, with four exceptions: (1) Correlations between accuracies for judging individual emotions or affective states within a test or between tests (e.g., anger correlated with surprise, joy correlated with disgust) were not included in order not to flood the database with inter-emotion accuracies. How accuracies for judging different specific emotions correlate with each other is undoubtedly an important research question, but one that exceeds the boundaries of the present project and that has been addressed elsewhere (e.g., Schlegel, Grandjean, & Scherer, 2012). (2) Correlations between a test total and subparts of the same test (i.e., part-whole correlations) were not included because item overlap would greatly inflate the correlations. (3) For the same reason, correlations between subparts of the same test were not included when participants could familiarize themselves with the stimuli (and the likely correct answers) due to exposure to the same stimuli in overlapping stimulus modalities; for example, the correlation between accuracy in judging the Multimodal Emotion Recognition Test's (MERT's; Bänziger, Grandjean, & Scherer, 2009) still pictures of facial expressions and accuracy at judging the MERT's video clips from which those still pictures were taken was not included because participants could learn about one from the other, thus inflating the correlation between the two accuracies. (4) If, within the same sample for a given test, correlations with other test(s) were reported for both the first test's total score and its subtest scores (e.g., PONS total score correlated with other tests and, in the same study, PONS face, body, and voice scores also correlated with the same other tests), only the effect sizes involving subtest scores were included

in the database. The respective total test score correlations were excluded to avoid multiple representations of the same data from which subtest and total scores had been calculated. Effect sizes based on subtest scores rather than total scores were chosen because this increased the variety in test characteristics such as cue channels and content domains in the database. (See a later section for a comparison of those excluded total scores' effect sizes to their respective subtests' effect sizes.) In reviewing studies for possible inclusion, we did not keep a record of how many were excluded according to specific exclusion criteria. However, by far the greatest number were excluded because they did not administer more than one IPA test; indeed, this was the rule and the exceptions were those relatively few that did include more than one test.

Final Database

All effect sizes were the Pearson correlation (*r*) between two IPA tests for a given sample of participants. We would have allowed cases where the relation between two tests might have been expressed as a partial correlation or a standardized regression coefficient, but none were identified. A given sample of participants would produce multiple effect sizes if more than two IPA tests, assuming the author reported their respective pairwise correlations. A given source (e.g., an article) could yield results for more than one independent sample of participants.

The database consisted of 83 sources containing 103 independent samples. In turn, those 103 samples yielded 622 effect sizes.

Coding of Source and Sample Characteristics

The descriptive statistics for these characteristics are provided in Table 1.

Year. This was the year of publication for a book or article, or the year provided by the author for unpublished results.

Publication status. The categories were: *journal article, thesis/dissertation, book, unpublished,* and "*published plus*," which meant that the original author provided between-test

correlations that were not included in the publication. Unpublished results that were published subsequent to retrieval were maintained as "unpublished" in the analysis, but the published citation to the work is given in the Reference section if it was known.

Size of participant sample on which the effect size was based.

Mean age of participant sample.

Percentage of females in participant sample.

Sample ethnicity. The categories were: *more than 60% White, more than 60% Black, more than 60% Hispanic, more than 60% other minority, multiple groups with none exceeding 60%*, and *multiple groups with no information on the proportions*.

Number of nested effect sizes. For each sample it was noted how many effect sizes were based on that sample, that is, nested within it. This variable was used to implement the nesting of effects within studies in multi-level analyses (see Statistical Analysis).

Coding of Test Characteristics

The descriptive statistics for these characteristics are provided in Table 2.

Content domain. This variable captured the state, trait, or attribute that participants were asked to judge. The categories were: *emotion*, judged either on a dimension (e.g., rate how happy the face is) or as a categorical choice (choose which emotion the face is expressing, from a multiple choice); *thoughts and feelings*, defined as in the empathic accuracy paradigm (Ickes, 2001), in which perceivers watch a person's spontaneous interpersonal behavior on video and guess what the person was thinking or feeling during or at the end of the clip, with the accuracy criterion being what the target person reported after viewing their own video (note, in keeping with the exclusion criteria, live dyadic studies using this paradigm were not included); *situational affect*, meaning participants selected what situation the target person was in, for example ordering food in a restaurant versus talking about her divorce; *deception*, defined as

whether target person was dissembling or telling the truth; *social attributes*, consisting of categorical variables describing (relatively) static social or group distinctions such as political allegiances or sexual orientation; and *personality traits,* for example conscientiousness or neuroticism. Table 3 lists specific tests and constructs falling into each of these content domain categories.

Cue channels. The cue channels in the test could include: *face; voice quality* (meaning the voice was audible but only its nonverbal qualities could be discerned, which would be accomplished through content masking [e.g., bandpass filtering] or by having targets recite ambiguous or meaningless linguistic content while varying intended messages such as different emotions); *body* (arms, legs, and/or torso, but not head); *linguistic* (coded when participants could understand the words the target persons were saying and the words contained potentially meaningful content); and *eyes* (coded only when stimuli consisted of only the eyes). Each of these five cue channels was coded as "yes" or "no," as applicable. Thus, if participants watched and listened to a video showing the whole person with the original audio track, the coder would check face, voice quality, body, and linguistic. A given test could thus consist solely of one of these cue channels, or any combination of them. Because the codes were applied to the test as a whole, no distinction was made in terms of what cues were combined within a specific test stimulus item. This means, for example, that a test where half the items showed only the face and half showed only the body would receive the same coding (face present, body present) as a test where all the items showed the whole person, that is, a body with its head attached.

Stimulus presentation mode. The stimulus material in the test could be: *static* (i.e., photographs); *dynamic* (films, videos, or vocal tracks); or *both*.

Stimulus creation mode. This coding variable consisted of: *posed* (expressions or behavior deliberately enacted for purposes of stimulus creation); *spontaneous* (expressions or behavior recorded under relatively unconstrained conditions, such as a "get acquainted" conversation, or during task performance when the target person was unaware of being observed); and *physical appearance* only.[1] This variable was not coded for studies on lie-detection accuracy because it was too ambiguous. The behavior of target persons telling the truth could be spontaneous (i.e., actually occurring, unscripted) but might also be posed in the sense that they are deliberately trying to seem honest. On the other hand, target persons who are lying might be posing (by trying to look honest), or they might be spontaneously behaving as they do when they lie.

Response format. Participants could respond to the stimuli in one of three ways: *forced choice* (i.e., multiple choice answer options); *dimensional ratings* (e.g., rating of how angry a face seemed), or *open response* (participants could write a free narrative of what they thought the answer was).

Standard test. Tests were classified as *standard* or *not standard*, where standard was defined as a named test that was used repeatedly in the literature and/or for which the developer had published at least one validity article. Variants, shortened versions, and subparts of such tests were equally considered standard. Authors of standard tests were likely to have explicitly addressed issues of reliability and validity and to have published normative data. Non-standard tests were, by contrast, typically developed for a particular study and were not known to be used by other investigators much, if at all. Although standard and non-standard tests were generally similar in their content and structure, the comparison of standard to non-standard was considered important because of the possibility that standard tests would be psychometrically superior and

therefore would have the capacity to correlate more strongly with other tests. Table 1 identifies tests classified as standard.

Reliability. *Internal consistency* (Cronbach's alpha or equivalent) for the given sample was recorded whenever authors provided it, and as an additional contributor to psychometric characteristics, *number of test stimuli* was also recorded.

Coder Accuracy

The authors each coded a subset of the studies. Accuracy of coding was confirmed in several ways. First, coding of sample characteristics was straightforward, as the information (e.g., sample size) was reported directly. Second, coding of test characteristics was confirmed by using templates for the most commonly occurring instruments (e.g., the full PONS test of Rosenthal et al., 1979, or the Diagnostic Analysis of Nonverbal Accuracy [DANVA-2] Adult Faces test of Nowicki & Duke, 1994). Also, the authors performed many cross-checks of each other's coding on a study-by-study basis and also within the database by internal checks to confirm that similar tests were coded in the same way. Third, calculation of effect sizes was never required, as correlations were provided directly by the original authors in their published works or in their personal communications about unpublished results.

Statistical Analysis

All analyses were conducted with IBM SPSS Statistics Version 22 (IBM Corp., 2013).

To account for the nesting of effect sizes within samples, the data were analyzed using multi-level modeling (MLM) with sample number as the random effects nesting variable. MLM has been proposed as a suitable framework for meta-analysis especially when effect size estimates are nested within studies and the number of effect sizes varies considerably between studies, as is the case here (e.g., Hox & De Leeuw, 2003; Konstantopoulos, 2011). There are

several published meta-analyses that used MLM to account for nesting (e.g., Acar & Sen, 2013; Allen, Chen, Willson, Hughes, 2009; Thompson & Voyer, 2014). Treating study or sample as a random effect in a mixed model takes into account two sources of sampling error – within studies and between studies – when estimating effect size. This analysis is analogous to a classical random effects meta-analysis, but additionally takes into account the possibility that effect sizes within one study are more similar than effect sizes across studies (Hox & De Leeuw, 2003). The random effects approach was chosen because we assumed that the "true" effect size could vary from study to study due to factors such as the reliability or nature of the specific tests or sample characteristics.

The dependent variable in all analyses was the Pearson correlation between a given pair of tests transformed into Fisher's z ($k = 622$ effect sizes altogether). The Fisher-z transformation normalizes values on the correlation scale. Effect sizes were transformed back to the *r*-metric for all data presentations. To assess the overall association between IPA tests, an unconditional means model with effect size as the dependent variable and no predictors was computed. For comparison, the average effect size without nesting was also computed. The estimation of random effects used the default "variance components" covariance structure in SPSS that assigns a scaled identity matrix to each specified effect. The degrees of freedom were estimated using the Satterthwaite correction, which approximates degrees of freedom when there are unequal variances and group sizes and yields fractional values.

To assess the moderating influence of test and sample characteristics on effect size, each potential moderator was separately added as a fixed effect to the unconditional means model (i.e., yielding random intercept models). Random-intercept models were chosen because the primary goal was to assess the effect of the different moderator variables on the average

correlation between two IPA tests (i.e., how the intercept varies as a function of each moderator variable), and to get an estimate of the correlation at each level of the target moderator. Random slopes that assess to what extent a moderator's effect varies between studies were not analyzed in these models as this question was not of immediate relevance for our goal. Further, slope parameters would be difficult to interpret for most (categorical) moderator variables in the present meta-analysis. Two types of moderators were analyzed.

Level 1 moderators. These varied on the effect size level and consisted of the following: Content domain combination of the two tests, cue channel combination, test type combination (standard or not), stimulus creation mode combination, stimulus presentation mode combination, and response format combination (all of these combination variables were categorical); and average reliability and average number of stimuli in the two tests (both variables were continuous). By "combination" we mean new variables that were constructed to describe similarity or difference between the two tests on the given characteristic. For instance, the combination variable for test type had the values "both tests standard," "one test standard, one test non-standard," and "both tests non-standard." The two continuous moderators were standardized prior to the analysis at the effect size level.

Due to the nature of the meta-analytic data, not all combinations of all the Level 1 moderators were available and the ones that were available often showed systematic overlap (i.e., domain and cue channel were not orthogonal). As such, and given the exploratory nature of this project, each Level 1 moderator was analyzed in a separate multi-level analysis.

Level 2 moderators. These varied on the sample level and consisted of the following: publication status, participant sample ethnicity (both categorical), publication year, percentage of

female participants, and mean age of participant sample (all continuous). Continuous moderators were standardized prior to the analysis at the sample level.

The effect of categorical moderators on effect size was assessed by adding each variable separately as a factor to the unconditional means model. The effect of continuous test characteristics was assessed by adding these variables separately as covariates to the unconditional means model.

Descriptive Statistics

Table 1 shows how often different IPA tests appeared in the meta-analysis, categorized by their content domain and whether they were standard or not. In total there were 135 unique tests, 76 of which measured accuracy in judging emotions. Table 2 gives a basic description of the 103 samples that provided the 622 effect sizes. Table 3 displays the frequencies of key test characteristics separately for each content domain. Table 3 reveals important differences in test characteristics according to content domain. For example, although tests of judging emotions and situational affect were nearly all based on posed stimuli, all of the personality judgment tests were based on spontaneous cues. Implications of the confounding between moderators will be discussed at a later point.

Results

Overall Correlation between Interpersonal Accuracy (IPA) Tests

The distribution of the 622 Pearson correlations (effect sizes) is shown in the stem-and-leaf display in Table 4. Correlations between tests ranged from r = -.43 to r = .85 with a mean correlation of r = .12 and a median of .11 (both uncontrolled for nesting). To estimate the global effect size controlling for the nesting of the 622 effect sizes within the 103 samples, we ran an unconditional means multi-level model without predictors. The average correlation between tests accounting for nesting as indicated by the intercept in this model was r = .19 (SE = .02; p <

.001). The estimated variance of the intercept was .02 and highly significant ($p < .001$), indicating that the different studies substantially varied in their effect sizes and can be considered heterogeneous. The intraclass correlation coefficient (ICC), calculated as the ratio of between study variance and total variance, was .43, suggesting that about 19% of the variance in effect sizes is explained by the study of origin.

The remaining sections explore structure within the IPA construct, as embodied in empirical correlations between tests, by examining variables that might influence how strongly tests are correlated with each other. Each Level 1 moderator was evaluated separately, thus there were no overall model statistics to report that included all the Level 1 moderators. The average correlation accounting for nesting represents the best measure of the overall effect.

Content Domain

In order to evaluate how the similarity or difference in the content domains assessed by the two correlated tests influenced effect sizes, we ran a multi-level model with the domain combination of the two tests as a Level 1 predictor. With each of the two tests measuring one of six content domains, there were 21 possible domain combinations, of which 19 occurred in the dataset. Table 5 shows the effect sizes estimated for each domain combination. Correlations were positive between all domains with the exception of *personality with deception*. For 11 of the 19 domain combinations, the mean effect size was significantly greater than zero. The largest correlation (mean $r = .38$, $p < .01$) was found when both tests measured accuracy in detecting deception (although it was based on only two effect sizes), and the smallest was found when one test measured accuracy in judging personality and the other accuracy in detecting deception (mean $r = -.03$, $p = .72$). The second largest correlation involved both tests measuring accuracy in judging emotions (mean $r = .29$, $p < .001$).

To explore which domain combinations statistically differed from each other, we conducted pairwise comparisons with Bonferroni-corrected *p*-values (*p*-values were multiplied by the number of comparisons; i.e., 19) as part of the multi-level model described above. The only domain combination that had a significantly larger average correlation than other combinations was *emotion with emotion* (meaning both tests measured emotions): The average correlation between two tests that both measured accuracy in judging emotions was significantly higher than the correlations for the combinations *thoughts and feelings with social attributes* ($p < .05$), *personality with personality* ($p < .001$), *social attributes with social attributes* ($p < .01$), *personality with emotion* ($p < .001$), *emotion with social attributes* ($p < .001$), *situational affect with social attributes* ($p < .05$), and *emotion with situational affect* ($p < .05$).

To further explore how the six content domains compared with respect to the other domains, for each of the six domains we ran one multi-level model with a dummy-coded variable (1 = effect size involved at least one test measuring this domain, 0 = effect size did not involve this domain) as a predictor. Each of these six analyses compared all effects involving one target domain against all effects not involving that same target domain. The analysis for personality showed that when correlations were based on at least one personality judgment test, they were significantly lower than all effect sizes not involving a personality test ($p < .01$). Similarly, the analysis for social attributes showed that when correlations were based on at least one social attributes test, they were significantly smaller than all effect sizes not involving social attributes ($p < .01$). When correlations were based on at least one emotion test, they were significantly larger than all other effect sizes that did not involve an emotion test ($p < .01$). For the other domains, the differences between effect sizes that did and did not include tests of the respective domain were not significant. Overall, the correlational pattern shown in Table 5 suggests that

tests of personality and social attributes were less highly correlated with tests of their same domain than tests of emotion, deception, or situational affect. Furthermore, situational affect, emotion, and deception were the domains that were most closely related. Finally, emotion emerged as a central domain, showing the highest correlations with other domains.

Cue Channels

In order to evaluate how the cue channel(s) assessed in each test influenced effect sizes, we ran a multi-level model with the similarity or difference in the cue channel combination of the two tests as a Level 1 predictor. For the individual tests, there were nine different cue channel configurations, which yielded 45 possible channel combinations between the two tests. The 29 of these possible combinations that occurred in the dataset constituted the categories of the cue channel combination predictor variable in the multi-level model. Table 6 shows the effect sizes estimated for each cue channel combination. Correlations ranged from r = -.08 (p = .62; for an effect size for *face with face and voice*—that is, a face test correlated with a face and voice test) to r = .38 (p = .06; for an effect size for *body with eyes*), although both combinations occurred only one single time in the dataset. Nineteen of the 29 combinations yielded correlations that were significantly above zero. Among the most frequently assessed combinations, the highest correlations were found for *face with face* (r = .28, p < .001) and for *face, voice, and body with face, voice, and body* (r = .25, p <. 001). Pairwise comparisons between all cue channel combinations with the Bonferroni-corrected p-values revealed that the correlation for *face with face* was significantly higher than the correlation for *face, voice, body, and linguistic cues with face, voice, body, and linguistic cues* (p < .01). No other cue channel combinations significantly differed from each other, which is partly related to the rather low frequencies in most combinations.

To explore further how the nine different channel configurations of the individual tests compared with respect to their average correlations, for each of the nine configurations we ran one multi-level model with a dummy-coded variable (1 = effect size involved at least one test assessing this channel configuration, 0 = effect size did not involve this channel configuration) as a predictor. Correlations involving at least one test assessing the *face, voice, body, and linguistic cues* were significantly lower than all other correlations ($p < .01$). For the other cue channel configurations, the differences were not significant. This finding is likely to be related to the fact that tests assessing *face, voice, body, and linguistic* cues mostly tested personality judgment and almost never emotion; and personality tests were less correlated with other domains than emotion tests. In line with the above-mentioned findings for the personality domain, tests assessing *face, voice, body, and linguistic* cues were also less highly correlated with other tests of the same configuration than tests assessing other configurations. Overall, the correlational pattern shown in Table 6 suggests that tests assessing two or three channels (unless they assess linguistic cues) overall tended to have the highest associations with other tests.

Test Reliability

Test reliability (internal consistency for the given sample as measured with Cronbach's α) and the average number of test stimuli of the two tests (which is a contributor to test reliability) were significantly positively related with effect size. Information on the reliability of *both* tests was available for only 74 of the 622 effect sizes and for these 74 cases the averaged reliability for the two tests ranged from .13 to .92 (mean α = .48, *SD* = .16). In Table 7, the intercept of .23 for the reliability analysis indicates that the correlation between two tests with an average mean test reliability (i.e., a combined α of .48) is expected to be $r = .23$. The estimate of .10 means that a one standard deviation change in the combined α of the two tests will lead to an

increase or decrease in the correlation between these tests of .10. For example, the correlation between two tests with a mean α of one standard deviation above the mean (i.e., .48 + .16= .64) would increase by .10 from r = .23 to r = .33.

Information on the number of stimuli for both tests was available for 610 of the 622 effect sizes and the average number of stimuli in the two tests ranged from 5 to 236 (M = 29.11, SD = 29.34). As above, the intercept of .18 suggests that the average correlation of two tests with a mean number of items of 29.11 will be r = .18. The estimate of .03 indicates that the correlation between two tests will change by .03 when their mean number of items increases or decreases by one standard deviation. That is, two tests with a mean number of items that is one standard deviation above the average (i.e., 29.11 + 29.34 = 58.45) would be expected to correlate at r = .21 (.18 + .03). Collectively these findings demonstrate that more reliable tests and longer tests tend to yield higher effect sizes.

To explore the implications of this finding further, we conducted an additional analysis on a subset of studies for which we were able to retrieve both correlations of total test scores (e.g., MERT total score) and the test's subtest scores (e.g., MERT audio, MERT still pictures, and MERT video subtests) with other tests. (Recall from Method that, for this subset, correlations of total test scores with other tests were not part of the dataset of k = 622 that was used for all analyses reported above.) Comparing correlations involving total test scores (which are based on longer and hence presumably more reliable item sets) to correlations involving subtest scores (which are based on shorter and presumably less reliable item subsets) allows for an additional test of the effects of test reliability on correlations between IPA measures. For nine of the samples in the present meta-analysis, both subtest and total score correlations of the same test with other measures were available. In these nine samples, there were 223 effect sizes (part

of the $k = 622$ dataset) that involved at least one subtest score of a test for which also total score correlations were available, and 58 effect sizes that involved at least one total score of a test for which also subtest score correlations were available (not part of the $k = 622$ dataset). We ran a multi-level model with the effect sizes as the dependent variable and the dummy coded variable 1 = "effect size involved at least one total test score" and 0 = "effect size involved only subtest scores" as a factor. The effect of this factor was significant, $F(2, 18.727) = 17.761$, $p < .001$, with the mean correlation for the 58 effect sizes involving total test scores being significantly higher than the mean correlation for the 223 effect sizes involving only subtest scores (.27 versus .20, difference significant at $p < .01$). This indicates that longer tests and full tests rather than subparts of tests generally yield higher correlations with other IPA measures, presumably through their higher reliability.

It is known that correlations between tests are restricted in size as a function of the amount of error variance in each test (Guilford, 1954). That is, the lower the reliabilities of the two tests are, the lower is the upper boundary for the correlation between them. Given that, overall, the internal consistencies of IPA tests tend to be low, the overall effect size of $r = .19$ that was found in the present meta-analysis might underestimate the relationship between the true variances of these tests. In order to explore how much test (un)reliability affected the correlations between tests, we corrected the effect sizes for attenuation using the Cronbach's alpha coefficients of both tests when they were available (Guilford, 1954) and estimated the overall effect size in a multi-level model. For the 74 effect sizes where Cronbach's alpha was available for both tests, the overall effect size was $r = .26$ without correction and $r = .40$ with correction for attenuation.

Other Moderators

Standard test. Whether the two tests were standard or not significantly influenced the correlation between two tests; effect sizes were significantly higher when both tests were standard (r = .23) than when one test (r = .14) or both tests (r = .13) were non-standard.

Stimulus presentation mode. Whether stimuli in a test were static (e.g., photos) or dynamic (e.g., videos, audio recordings) did not have a significant effect on the correlation between two tests.

Stimulus creation mode. Stimulus creation mode was a significant moderator of effect sizes. When both tests used posed stimuli, correlations were significantly stronger (r = .25) than when one test (r = .15) or both tests (r = .11) used stimuli showing spontaneous target behavior or when both tests were based on physical appearance (r = .09).

Response format. Response format was significantly associated with the correlation between two tests. When both tests used a forced-choice response format, correlations were significantly stronger (mean r = .23) than when one test (r = .07) or both tests (r = .08) used dimensional rating scales.

It needs to be pointed out that response format, stimulus creation mode, and test standardization were highly confounded with content domain in our dataset. As was seen in Table 3, standard tests, posed stimuli, and forced choice response format are prevalent mostly in the emotion and situational affect domains that had yielded higher correlations between tests than, for example, the personality domain. Due to the low frequency of these features in tests of other domains, the effects of study characteristics cannot be disentangled from content domain effects.

Level 2 Moderators: Source and sample characteristics. The effect of publication status and ethnic composition on the overall effect size was assessed by adding each of these

categorical variables separately as a factor to the unconditional means model. The effect of continuous sample characteristics (sample size, percentage of female participants, publication year, and mean age) was assessed by adding these variables separately as covariates to the unconditional means model, resulting in four multi-level models. The results are presented in Table 7. None of the source or sample characteristics significantly affected the correlation between two tests.

Publication Bias

In the present meta-analysis, nearly half of all effect sizes were obtained from unpublished studies or upon request from authors of published studies who had not included the relevant results in their publication (see Table 2). Publication status was analyzed as a moderator and the effect was not significant (see Table 7), showing that effect sizes of published studies were not larger than effect sizes of unpublished studies. As recommended by Hox and De Leeuw (2003), we also analyzed sample size as a moderator in order to examine whether large positive effect sizes were predominantly found in smaller studies, which could suggest that many nonsignificant or negative effects obtained in other small studies might have remained unpublished. In our analysis, sample size was unrelated to effect size (see Table 7), speaking against this potential bias. In order to assess the number of studies with a zero effect size needed to bring the overall unnested mean effect of $r = .12$ to a trivial magnitude (we chose $r = .05$), we calculated Orwin's fail-safe N (Orwin, 1983), using the Comprehensive Meta-Analysis software (Borenstein, Hedges, Higgins, & Rothstein, 2005). By this calculation, it would take 795 additional studies with mean correlation of zero to bring the combined correlation under $r = .05$. In our view, this is an implausible number of unretrieved studies. Furthermore, if they do exist they only confirm our overall conclusion that the relation between tests is very small on average.

Discussion

Research on accuracy of interpersonal perception has a very long history (e.g., Buzby, 1924) and it is currently burgeoning (Hall et al., 2016). There are many traditions within this field, and many different kinds of questions that can be asked about interpersonal accuracy (IPA). Here, we focused on whether this ability is comprised of one skill or many skills. Zebrowitz (2001), discussing various approaches to measuring IPA, likened researchers in this field to the blind men who all thought they had touched a different animal because they had touched different parts of an elephant. Although there is likely much truth to this analogy, it may be equally likely that many IPA researchers are doing the opposite: touching different animals while thinking they are touching the same one—in other words, measuring different kinds of IPA but assuming they are measuring the same thing. The analyses presented here shed light on which interpretation is the more accurate representation of IPA structure, at least as it is revealed in the empirical data available to date: Is there, first, one general skill in perceiving others regardless of how IPA is measured; second, many discrete and unrelated skills (as many as there are instruments to measure skill); or third, a set of correlated skills each of which refers to a specific domain or type of interpersonal judgment? We answered this question by examining how different IPA tests correlate with each other and what factors might account for differences in strength among the various correlations.

The multi-level meta-analysis demonstrated a modest overall between-tests correlation of r = .19 that varied as a function of a number of different moderators. IPA tests within content domains were generally more highly related than IPA tests across domains. Some domains, in particular emotion recognition, appeared to be more homogeneous than other domains, meaning that correlations within this domain were higher than the average between-tests correlation of r = .19. Emotion recognition tests also yielded the highest correlations with other domains, whereas

personality and social attributes tests yielded the lowest correlations. With respect to cue channels, correlations tended to be higher the more nonverbal cue channels (face, voice, and body) were contained in the stimuli. Standard tests and similarly constructed tests also showed higher correlations, specifically, when both tests used posed stimuli and a forced-choice response format.

Finally, analyses of test length and reliability suggest that the average effect size increases if the low internal consistency of many tests is accounted for. When we recalculated overall effect size only with data from full tests rather than data from subtests, the overall correlation between tests improved significantly from $r = .20$ to $r = .27$. Furthermore, in an analysis based on correlations that were corrected for unreliability of the two tests, the average correlation between tests increased substantially from $r = .26$ to $r = .40$, indicating that the true association between people's skills in recognizing various traits, states, and attributes across different tests is not trivial, although one must remember that in practice no IPA tests have perfect reliability.

Mapping the Terrain of IPA

Given the significant positive overall correlation between IPA tests, the possibility that IPA is a collection of unrelated and instrument-specific skills can be dismissed. However, the overall effect size, even when accounting for test unreliability, was rather modest when compared to the strong correlations between tests in other psychological domains such as cognitive ability or Big Five personality dimensions (Dodrill, 1981; DeYoung, Quilty, & Peterson, 2007). Hence, the possibility that all tests measure the same global ability can be also be dismissed.

The present pattern most closely supports the hierarchical perspective as described in the third theoretical possibility. It suggests a set of domain- and channel-specific skills that are connected by a more general IPA (as reflected by the positive overall correlation between domains), with each of the tests being connected to the respective domain- and channel-specific skills (as reflected by higher correlations among tests of the same domain than among tests of different domains). In other words, the structure that emerged from the meta-analysis suggests there are *kinds of accuracy*, not simply *different tests,* and that these kinds of accuracy are connected by a higher-order global IPA skill. Moderator analyses showed that such *kinds or facets* of accuracy can be represented by different content domains as well as certain test characteristics. Overall, tests that were similar in content and/or test characteristics were more highly correlated with each other. The impact of test similarity on the correlations between tests is well illustrated by examining correlations that were particularly strong. Eighteen effect sizes were larger than $r = .50$. For 12 of these (67%), both tests measured emotion judgment; if situational affect and thoughts/feelings are added in (they are, in a looser sense, about emotion), this figure jumps to 17 out of 18 (94%).

From a theoretical psychometric perspective, these results are consistent with the *"causal indicators"* model of measuring theoretical constructs proposed by Bollen and Lennox (1991). This model proposes that theoretical constructs can only be measured in a comprehensive and valid way when all facets that causally determine the construct are assessed, e.g., by different tests. Critically, the less these facets or "causal indicators" are correlated, the more unique incremental variance they can each add to explaining the global theoretical construct. The different facets are therefore not interchangeable, and omitting any of the facets would omit part of the overall construct. Bollen and Lennox (1991) contrasted this *"causal indicators"* model

with the conventional *"effect indicators"* model in which measures of a construct are not seen as its potentially independent facets, but as "effects" caused by the construct. In this model, each indicator represents a repeated effect or outcome of the same construct and the indicators are supposed to be substantially correlated (which is why most test developers hope their test items are strongly positively correlated with each other). The results of the present meta-analysis suggest that the *"causal indicators"* approach can appropriately describe the IPA domain: the variety of weakly correlated IPA measures across domains and channels captures many different facets of global IPA and collectively adds to its construct validity. Tests that measure similar content or use similar methodologies can be seen as clusters of causal indicators, among which the different tests are still not interchangeable (Bänziger, Scherer, Hall, & Rosenthal, 2011).

Methodological and Practical Implications

Although the issue of correlations among IPA tests and whether there is a sensible structure to the IPA domain may seem like a question of only methodological or psychometric interest, in fact it touches upon a number of important and timely questions facing the interpersonal accuracy field. Aside from helping to resolve a long-standing puzzle—brought about by previous authors' observation that tests of IPA do not correlate with each other very well (Buck, 1984; Colvin & Bundick, 2001; Hall, 2001)—the present results can help guide future research in substantive ways. In the remainder of the discussion, we will address practical implications for users and developers of IPA tests, theoretical implications for the understanding of mechanisms and processes underlying individual differences in IPA, and directions for future research in the IPA field.

Choosing an IPA measure. The present results clearly showed that IPA tests are not interchangeable. Although the average effect size was highly significantly above zero, any two

tests plucked at random from all of the 135 tests that were used in the database would share little variance with each other. Importantly, low internal consistency only partly accounts for this: Even error-free estimated between-test correlations were only of moderate magnitude, meaning that different tests (or kinds of tests) would still not be fully interchangeable even if they all had perfect reliability.

However, based on our own close familiarity with the field, we can say with confidence that many authors of studies on IPA, including ourselves at times, often give little to no justification for their choice of a particular IPA test. Certainly, domain of interest often dictates from which subset of tests a researcher is likely to choose. However, beyond that, due to their familiarity with (or loyalty to) one test more than others, or convenience (ease of administration and scoring, time required), or simple emulation of previous studies, the choice of a content domain of IPA, and of a given test within a chosen content domain, often seems arbitrary and certainly not explicitly justified either empirically or theoretically. On the one hand, a major reason for this is the absence of empirical data and prior theoretical reasoning to guide researchers in the choice of domains and tests. On the other hand, a large contributor to the lack of guidance in the literature is the fact that, either for pragmatic reasons or oversight, authors rarely measure IPA using more than one test. In searching the literature for the present meta-analysis, only a tiny fraction of all the available studies on IPA used more than one test. This fact makes it not only far more difficult to answer questions about the structure of IPA, but also retards the development of knowledge about the differential predictive validity of IPA tests and the constructs they represent. This latter issue goes far beyond the simple question of how tests correlate with each other but depends, in part, on knowing the answer to that question.

Predictive validity of different IPA facets. Collectively, there is ample evidence that IPA tests have significant predictive validity for a wide range of external variables (for reviews, see Davis & Kraus, 1997; Hall et al., 2009, 2016; Rosenthal et al., 1979). However, given the current findings that imply only partial overlap across measures of IPA, it is important to know whether this predictive validity is a broader aspect of social functioning related to the shared accuracy across IPA or is more specific to the domain in which it was tested. This question also leads to some other intriguing possibilities in terms of where there may be gaps in the literature and in terms of what types/classes of outcomes investigators who wish to compare the differential predictive validity of various tests might study. One example of a gap in understanding predictive validity can be found in a recent meta-analysis on IPA in relation to perceivers' power, status, and dominance (Hall et al., 2015). Those authors found that so many studies measured IPA in terms of affect recognition that content domain could not be examined as a moderator.

As an example of the expansion of knowledge that would result if researchers did administer more than one IPA test, consider a researcher who wants to find out what skills are most valuable for personnel recruiters to have, in terms of the success of their hiring recommendations. Should the recruiters have special skill in judging applicants' emotions, in judging their personality, or in judging whether they are telling the truth or not? Or are all three skills equally valuable?

Psychometric issues. The internal consistency or reliability of all the measures included in the present meta-analysis is likely to have played an important role in understanding the nature of the shared qualities of IPA. Unfortunately, the analysis was greatly impacted by the under-reporting of reliability coefficients throughout the literature, which is particularly problematic

because the reliabilities that were reported suggested that there was room for concern. For studies that did report reliability (Cronbach's alpha) for both tests, the average was .48, a figure that falls below conventional standards for good psychometric quality (Schmitt, 1996). Several researchers have previously noted that weak internal consistency is a likely reason for low correlations between IPA tests (Hall et al., 2005; Kenny, 2013; Davis & Kraus, 1997), given that internal consistency sets a boundary for the maximal correlation between two measures due to measurement error (Schmitt, 1996).

Several observations can be made to help researchers think about the issue of internal consistency in IPA tests. First, some IPA tests have very good reliability, for example the GERT, an emotion recognition test with Cronbach's alpha around .80 (Schlegel, Grandjean, & Scherer, 2014; Schlegel & Scherer, 2015). It is worth noting that the correlations involving the GERT were among the highest effect sizes in the present database. Second, there is ample evidence that individual differences in IPA exist, as demonstrated by the many studies showing predictive validity such as reviewed elsewhere in this article. Also, re-test reliability—when examined—has often been good (Hall et al., 2005). Despite previous evidence that individual differences are hard to detect for lie detection (Aamodt & Custer, 2006; Bond & DePaulo, 2008), the present meta-analysis found substantial correlations in the few cases where lie detection tests were correlated with each other, suggesting there are individual differences. Lie detection research might be greatly enhanced by the inclusion of more than one test of lie detection ability or by including other measures of IPA. In particular, tests of emotion recognition paired with lie detection in which the lie (or truth) is detected through emotional cues could yield particularly compelling results. Regardless, given evidence of predictive validity, re-test reliability, and individual differences, perhaps we should not be very concerned with internal consistency. On

the other hand, if tests had better internal consistency, validity coefficients would, logically, be stronger.

Third, consistent with the logic of the "causal indicators" model of measurement introduced earlier (Bollen & Lennox, 1991), one could argue that a test with weak internal consistency in fact gains conceptual strength by measuring several loosely related facets of the construct and embracing a wider universe of stimuli (e.g., items tapping into different cue channels or content) rather than many tightly linked constructs (e.g., all items in the same domain, same cue channel, etc.). Some tests, in fact, have a demonstrated factor structure (PONS, analysis in Rosenthal et al., 1979; MERT, analysis in Schlegel, Grandjean, & Scherer, 2012), and some were designed a priori to measure more than one kind of content (IPT; Costanzo & Archer, 1989).

Another important methodology-related finding in the present meta-analysis was that test characteristics such as response format and stimulus creation mode were largely specific to certain content domains and were not evenly distributed across domains. For example, tests in the emotion and situational affect domains mostly used forced choice response formats and consisted of stimuli in which the targets had been instructed to pose an emotion or affective state. Accordingly, these tests are typically scored by determining whether a judge chose the emotion label or affective state that corresponds to what the target had been asked to express. In contrast, tests about judgments of others' personality always used dimensional ratings of targets' spontaneous behavior and are typically scored with respect to the target's self- or informant-reported trait level. The current status of the field where the most commonly used tests and similar test construction practices are aligned with specific domains of IPA makes it difficult to reach a conclusion about whether variations in the degree of correlation between IPA tests were

caused by the domain or by test features. Our moderator analysis showed increased correlations between tests that both used a forced-choice response format which could simultaneously explain the higher correlation between two emotion recognition tests over a test of emotion recognition with a personality judgment task or two personality judgment tasks.

Another aspect that might have interacted with content domain in affecting between-test correlations was whether an IPA test was a standard instrument or not. Notably, almost half of the individual measurements in the dataset were based on non-standard tests. Standard IPA tests generally yielded higher correlations with other tests, but at the same time standard tests mostly measured emotion recognition, whereas there was not a single standard test measuring accuracy in judging personality. The low intercorrelations of personality judgment tests might or might not be partly related to the lack of standardized and potentially better validated tests in this domain.

As a direction for the psychometric future of the IPA field, it would be desirable to develop more standard tests, especially in the personality domain. The development of new tests could benefit from modern psychometric methods such as Item Response Theory (IRT; Embretson & Reise, 2000) and factor analysis (Brown, 2006). The use of these methods provides advanced information about a test's internal structure and consistency (such as whether the test's structure remains the same across cultures, genders, etc.), which can in turn inform the understanding of the IPA construct. It would also be desirable to develop tests that measure several domains simultaneously and/ or that include different stimulus channels and item types.

Development, mechanisms, and determinants of IPA. While the meta-analysis definitively shows support for some degree of relationship between the different measures of IPA, it does not directly inform about what the shared qualities of IPA might be. Having considered whether there is a shared construct that spans IPA, several more questions

immediately follow; for example: What are the mechanisms underlying accurate interpersonal judgments in the different domains, and what mechanisms and processes are shared between the domains? One result of the present analysis was that emotion recognition appeared to play a central role among all IPA domains. The processing of affective stimuli might thus be a shared feature of some, though perhaps not all, types of interpersonal judgments. This idea is supported by a range of neuroscientific studies that identified and compared the brain regions involved in different IPA and social cognition tasks (e.g., Heberlein & Saxe, 2005; van Overwalle, 2009). Another recent line of research examines *embodiment* as a mechanism in different types of interpersonal judgments (Niedenthal, Barsalou, Winkielman, Krauth-Gruber, & Ric, 2005). Embodiment describes the perceptual, somato-visceral, and motoric re-experiencing or simulation of a target person's state or behavior in the observer. Although to date there is no agreement yet about the precise nature of the mechanisms involved in making various interpersonal judgments, research in fields like neuroscience is likely to contribute to the understanding of IPA in the future.

Finally, yet another set of questions that emerge from this meta-analysis is which factors might influence whether people get specialized (or not) in one or another domain or cue channel. One variable that is likely to be related to higher overall IPA is general mental ability, including indicators of better information processing as well as better crystallized knowledge, for example in vocabulary (Murphy & Hall, 2011). With respect to specialization, in looking across the different literatures, there is evidence that people can be trained to become more accurate in a variety of IPA tasks (Blanch-Hartigan et al., 2012; Matsumoto & Hwang, 2011). Alternatively, some individuals have life experiences that may make them better, as suggested for example by research that shows that criminals may be better in lie detection than other groups (Hartwig,

Granhag, Strömwall, & Andersson, 2004; Vrij & Semin, 1996). There are also variables that could be related to being more accurate in specific cue channels, such as individual differences in global versus local processing of visual stimuli (Martin, Slessor, Allen, Phillips, & Darling, 2012).

Summary and Outlook

Early research attempting to identify a global ability of IPA produced lackluster results that left some researchers in the field wondering if there was any connection between the various tests and measures that have been used to assess IPA. The present meta-analysis represents a significant improvement over previous efforts to look for relationships across a large number of tests. Results showed that there is shared variance across content domains and that the more similar the methodologies and cue channels used in test construction, the higher the degree of positive relationship between measures. Also, there was some distinctiveness between the various content domains of IPA, suggesting that the various measures of IPA are not interchangeable.

There are also opportunities to explore new domains of interpersonal judgments that have not yet been well-connected to other forms of IPA. Such domains might include the accurate perception of others' pain (e. g. Ruben & Hall, 2013), the accurate appraisal of another person's emotion regulation skills (Murphy, 2014), or the ability to make accurate judgments about how others perceive oneself (meta-perception accuracy; Carlson & Barranti, 2016). Given the expanding interest in IPA, it is hoped that the present meta-analysis will serve as a starting point for future efforts to understand and measure this critical component of social competence.

References

Note: Citations (i.e., sources) marked with * contributed data to the meta-analysis. The phrase "counted with" means that the study sample was the same as used in another source and the data were ascribed to the other source.

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *The Forensic Examiner, 15*, 6-11.

Acar, S., & Sen, S. (2013). A multilevel meta-analysis of the relationship between creativity and schizotypy. *Psychology of Aesthetics, Creativity, and the Arts, 7*(3), 214–228.

Adams, H. F. (1927). The good judge of personality. *Journal of Abnormal Psychology, 22*, 172-181.

*Alaerts, K., Nackaerts, E., Meyns, P., Swinnen, S. P., & Wenderoth, N. (2011). Action and emotion recognition from point light displays: An investigation of gender differences. *PLoS ONE, 6*, e20989.

Allen, C. S., Chen, Q., Willson, V. L., & Hughes, J. N. (2009). Quality of research design moderates effects of grade retention on achievement: A meta-analytic, multi-level analysis. *Educational Evaluation and Policy Analysis, 31*(4), 480–499.

Ambady, N., Hallahan, M., & Conner, B. (1999). Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of Personality and Social Psychology, 77*, 538-547.

*Ambady, N., Hallahan, M., & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology, 69*, 518-529.

Ambady, N., & Skowronski, J. J. (Eds.) (2008). *First impressions*. New York: Guilford Publications.

App, B., McIntosh, D. N., Reed, C. L. & Hertenstein, M. J. (2011). Nonverbal channel use in communication of emotion: How may depend on why. *Emotion, 11*, 603-617.

Archer, D., & Akert, R. (1977). Words and everything else: Verbal and nonverbal cues in social interaction. *Journal of Personality and Social Psychology, 35*, 443-449.

*Back, M. D., Schmukle, S. C. & Egloff, B. (2008). Becoming friends by chance. *Psychological Science, 19*, 439-440.

*Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in the face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion, 9*, 691-704.

Bänziger, T., Scherer, K. R., Hall, J. A., & Rosenthal, R. (2011). Introducing the MiniPONS: A short multichannel version of the Profile of Nonverbal Sensitivity (PONS). *Journal of Nonverbal Behavior, 35,* 189-204.

*Barnes, M. L., & Sternberg, R. J. (1989). Social intelligence and decoding of nonverbal cues. *Intelligence, 13*, 263-287.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*, 241-247.

*Baum, K. M., & Nowicki, S., Jr. (1998). Perception of emotion: Measuring decoding accuracy of adult prosodic cues varying in intensity. *Journal of Nonverbal Behavior, 22*, 89-107.

Bernieri, F. J. (1991). Interpersonal sensitivity in teaching interactions. *Personality and Social Psychology Bulletin, 17*, 98-103.

*Bernieri, F. J., & Gillis, J. S. (1995). Personality correlates of accuracy in a social perception task. *Perceptual and Motor Skills, 81*, 168-170.

*Bernieri, F. J. (2012). Unpublished data, Oregon State University.

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., et al. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21*, 3-21.

*Blanch, D. C., & Andrzejewski, S. A. (2009). Unpublished data, Northeastern University.

Blanch-Hartigan, D., Andrzejewski, S. A., & Hill, K. M. (2012). The effectiveness of training to improve person perception accuracy: A meta-analysis. *Basic and Applied Social Psychology, 34*, 483-498.

*Blanch-Hartigan, D. (2011). Measuring providers' verbal and nonverbal emotion recognition ability: Reliability and validity of the Patient Emotion Cue Test (PECT). *Patient Education and Counseling, 82*, 370-376.

*Blanch-Hartigan, D. (2012). An effective training to increase accurate recognition of patient emotion cues. *Patient Education and Counseling, 89*, 274-280.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305–314.

Bond, C. F., Jr., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134*, 477-492.

Boone R. T, & Buck R. (2003). Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior, 27*, 163-182.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis* (2nd ed.). Englewood, NJ: Biostat.

Borkenau, P., & Liebler, A. (1992). The cross-modal consistency of personality: Inferring strangers' traits from visual or acoustic information. *Journal of Research in Personality, 262*, 183-204.

*Bowman, J. K. (2006). *The utility of emotional intelligence as a predictor of school psychologists clinical competence*. Unpublished dissertation, St. John's University.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford Press.

Brüne, M., & Schaub, D. (2012). Mental state attribution in schizophrenia: What distinguishes patients with "poor" from patients with "fair" mentalising skills? *European Psychiatry, 27*, 358-364.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

Buck, R. (1976). A test of nonverbal receiving ability: Preliminary studies. *Human Communication Research, 2*, 162-171.

*Buck, R. (1979). Measuring individual differences in the nonverbal communication of affect: The slide-viewing paradigm. *Human Communication Research, 6*, 47-57.

Buck, R. (1984). *The communication of emotion*. New York: Guilford Press.

Buzby, D. E. (1924). The interpretation of facial expression. *American Journal of Psychology, 35*, 602-604.

Byron, K. (2008). Differential effects of male and female managers' non-verbal emotional skills on employees' ratings. *Journal of Managerial Psychology, 23*, 118-134.

Carlson, E. N., & Barranti, M. (2016). The accuracy of metaperceptions: Do people know how others perceive them? In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social*

psychology of perceiving others accurately (pp. 165-182). Cambridge, UK: Cambridge University Press.

*Carney, D. R. (2002). Unpublished data, Northeastern University.

*Carney, D. R. (2009). Unpublished data, Columbia University.

*Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*, 1054-1072.

*Carton, J. S., Kessler, E. A., & Pape, C. L. (1999). Nonverbal decoding skills and relationship well-being in adults. *Journal of Nonverbal Behavior, 23*, 91-100.

*Carton, J., & Nowicki, S., Jr. (1993). The measurement of emotional intensity from facial expressions. *Journal of Social Psychology, 133*, 749-750.

*Castro, V., & Boone, R. T. (2015). Sensitivity to spatiotemporal percepts predicts the perception of emotion. *Journal of Nonverbal Behavior*. (no pagination available yet)

Cline, V. B., & Richards, J. M., Jr. (1960). Accuracy of interpersonal perception—A general trait? *Journal of Abnormal and Social Psychology, 60*, 1-7.

*Colvin, C. R., & Bundick, M. J. (2001). In search of the good judge of personality: Some methodological and theoretical concerns. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 47-65). Mahwah, NJ: Lawrence Erlbaum Associates.

*Costanzo, M., & Archer, D. (1989). Interpreting the expressive behavior of others: The Interpersonal Perception Task. *Journal of Nonverbal Behavior, 13*, 225-245.

Crow, W. J., & Hammond, K. R. (1957). The generality of accuracy and response sets in interpersonal perception. *Journal of Abnormal and Social Psychology, 54*, 384-390.

*Davies, M., Stankov, L., & Roberts, R. D. (1998). Emotional intelligence: In search of an elusive construct. *Journal of Personality and Social Psychology, 75*, 989-1015.

Davis, M. H., & Kraus, L. A. (1997). Personality and empathic accuracy. In W. J. Ickes (Ed.), *Empathic accuracy* (pp. 144–168). New York: Guilford Press.

DeMeijer, M. (1989). The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, *13*, 247-268.

*Demenescu, L. R., Mathiak, K. A., & Mathiak, K. (2014). Age- and gender-related variations of emotion recognition in pseudowords and faces. *Experimental Aging Research, 40*, 187-207.

DePaulo, B. M., LeMay, C. S., & Epstein, J. A. (1991). Effects of importance of success and expectations for success on effectiveness at deceiving. *Personality and Social Psychology Bulletin, 17*, 14-24.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93*, 880-896.

Dodrill, C. B. (1981). An economical method for the evaluation of general intelligence in adults. *Journal of Consulting and Clinical Psychology, 49*, 668-673.

Ekman, P. (2002). *MicroExpression training tool (METT)*. San Francisco: University of California.

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Palo Alto: Consulting Psychologists Press.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A technique for the measurement of facial movement.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., & Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology, 53*, 712-717.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 46*(9), 913-920.

Elfenbein, H. A., Foo, M. D., White, J., Tan, H. H., & Aik, V. C. (2007). Reading your counterpart: The benefit of emotion recognition accuracy for effectiveness in negotiation. *Journal of Nonverbal Behavior, 31*(4), 205–223.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Hillsdale, NJ: Erlbaum.

*Feldman, M., & Thayer, S. (1980). A comparison of three measures of nonverbal decoding ability. *Journal of Social Psychology, 112*, 91-97.

Feleky, A. M. (1914). The expression of the emotions. *Psychological Review, 21*, 33-41.

*Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology, 72*, 1429-1439.

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55*, 149-158.

Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. Journal of Personality and Social Psychology, 64, 479-490.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Cultural relativity in perceiving emotion from vocalizations. *Psychological Science, 25*, 911-920.

*Gesn, P. R., Bernieri, F. J., Grahe, J. E., & Gada-Jain, N. (1999, April). *Domains of interpersonal sensitivity: Performance accuracy and self-reports of ability*. Presented at Midwestern Psychological Association, Chicago. (Cited in Hall, J. A., (2001). The PONS test and the psychometric approach to measuring interpersonal sensitivity. In J. A. Hall & F. J. Bernieri (Eds.). (2001). *Interpersonal sensitivity: Theory and measurement* (pp. 143-160). Mahwah, NJ: Lawrence Erlbaum Associates.).

*Goldstein, T. R. (2011). Correlations among social-cognitive skills in adolescents involved in acting or arts classes. *Mind, Brain, and Education, 5*, 97-103.

*Golombeck, N. (2005). *Attachment style and nonverbal communication*. Unpublished dissertation, St. John's University.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Halberstadt, A. G. (1986). Family socialization of emotional expression and nonverbal communication styles and skills. *Journal of Personality and Social Psychology, 51*, 827-836.

Hall, J. A. (2001). The PONS test and the psychometric approach to measuring interpersonal sensitivity. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity*: *Theory and measurement* (pp. 143-160). Mahwah, NJ: Erlbaum.

Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior, 33,* 149-180.

Hall, J. A., Bernieri, F. J., & Carney, D. R. (2005). Nonverbal behavior and interpersonal sensitivity. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 237-281). Oxford: Oxford University Press.

*Hall, J. A., & Carter, J. D. (1999). Gender-stereotype accuracy as an individual difference. *Journal of Personality and Social Psychology, 77*, 350-359.

*Hall, J. A., & Goh, J. X. (2014). Unpublished data, Northeastern University.

*Hall, J. A., & Gunnery, S. D. (2014). Unpublished data, Northeastern University.

Hall, J. A., Gunnery, S. D., & Horgan, T. G. (2016). Gender differences in interpersonal accuracy. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 309-327). Cambridge, UK: Cambridge University Press.

*Hall, J. A., Halberstadt, A. G., & O'Brien, C. E. (1997). "Subordination" and nonverbal sensitivity: A study and synthesis of findings based on trait measures. *Sex Roles, 37*, 295-317.

*Hall, J. A., Roter, D. L., Blanch, D. C., & Frankel, R. M. (2009). Nonverbal sensitivity in medical students: Implications for clinical interactions. *Journal of General Internal Medicine, 24*, 1217-1222.

*Hall, J. A., Ruben, M. A., & Curtin, L. (2012). Unpublished data, Northeastern University.

Hall, J. A., Schmid Mast, M., & Latu, I. (2015). The vertical dimension of social relations and accurate interpersonal perception: A meta-analysis. *Journal of Nonverbal Behavior, 39,* 131-163.

Hall, J. A., Schmid Mast, M., & West, T. V. (Eds.) (2016). *The social psychology of perceiving others accurately*. Cambridge, UK: Cambridge University Press.

*Hall, J. A., Ship, A. N., Ruben, M. A., Curtin, E. M., Roter, D. L., Clever, S. L., Smith, C. C., & Pounds, K. (2014). The Test of Accurate Perception of Patients' Affect (TAPPA): An ecologically valid tool for assessing interpersonal perception accuracy in clinicians. *Patient Education and Counseling, 94,* 218-223.

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Andersson, L. O. (2004). Suspicious minds: Criminals' ability to detect deception. *Psychology, Crime & Law, 10*, 83-95.

Hodgins, H. S., & Koestner, R. (1993). The origins of nonverbal sensitivity. *Personality and Social Psychology Bulletin, 19*, 466-473.

Hox, J. J., & de Leeuw, E. D. (2003). Multilevel models for meta-analysis. In S. P. Reise & Duan, N. (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 90–111). Mahwah, NJ: Lawrence Erlbaum Associates.

Orwin, R. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*, 157–159.

IBM Corporation (2013). *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp.

Ickes, W. (2001). Measuring empathic accuracy. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 219–241). Mahwah, NJ: Lawrence Erlbaum Associates.

Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin, 115*, 288-299.

*Jabeen, L. N., Boone, R. T., & Buck, R. (2009). Unpublished data, University of Massachusetts, Dartmouth.

Kenny, D. A. (2013). Issues in the measurement of judgmental accuracy. In S. Baron-Cohen, H. Tager-Flusberg, & M. V. Lombardo (Eds.), *Understanding Other Minds: Perspectives from Developmental Social Neuroscience* (pp. 104–116). Oxford, UK: Oxford University Press.

Kurkul, W. W. (2007). Nonverbal communication in one-to-one music performance instruction. *Psychology of Music, 35*, 327-362.

*Klaiman, S. (1979). *Selected perceptual, cognitive, personality, and socialization variables as predictors of nonverbal sensitivity*. Unpublished doctoral dissertation, University of Ottawa. (counted in Buck, 1979)

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2*(1), 61–76.

*Krendl, A. C., Rule, N. O., & Ambady, N. (in press). Does aging impair first impression accuracy? Differentiating emotion recognition from complex social inferences. *Psychology and Aging*.

*Lambrecht, L., Kreifelts, B., & Wildgruber, D. (2012). Age-related decrease in recognition of emotional facial and prosodic expressions. *Emotion, 12*, 529-539.

*Lewis, K. L., & Hodges, S. D. (2009). Unpublished data, University of Oregon.

*Lippa, R. A., & Dietz, J. K. (2000). The relation of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior, 24*, 25-43.

*Littlepage, G. E., McKinnie, R., & Pineault, M. A. (1983). Relationship between nonverbal sensitivities and detection of deception. *Perceptual and Motor Skills, 57*, 651-657.

*Locher, B., Lewis, K. L., & Hodges, S. D. (2009). Unpublished data, University of Oregon.

Martin, D., Slessor, G., Allen, R., Phillips, L. H., & Darling, S. (2012). Processing orientation and emotion recognition. *Emotion*, *12*, 39–43.

Matsumoto, D., & Hwang, H. S. (2011). Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion, 35*, 181-191.

Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., et al. (2000). A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). *Journal of Nonverbal Behavior, 24,* 179-209.

Mayer, J. D., Panter, A. T., & Caruso, D. R. (2012). Does personal intelligence exist? Evidence from a new ability-based measure. *Journal of Personality Assessment, 94*(2), 124–140.

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. Sluyter (Eds.), *Emotional development and emotional intelligence: Educational implications* (pp. 3–31). New York: Basic Books.

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3*, 97-105.

*McIntire, K. A., Danforth, M. M., & Schneider, H. G. (1999). Measuring cue perception: Assessment of reliability and validity. *North American Journal of Psychology, 1*, 261-266.

*Mill, A., Allik, J., Realo, A., & Valk, R. (2009). Age-related differences in emotion recognition ability: A cross-sectional study. *Emotion, 9*, 619-630.

*Mortimer, D. C. (1976). Unpublished master's thesis, University of Texas at Arlington. Cited in W. Ickes (2003), *Everyday mind reading*, Amherst, NY: Prometheus Books.

Murphy, N. A. (2014, May). *Emotion regulation and social interactions*. Paper presented at the Perspectives on Interpersonal Accuracy Workshop, University of Neuchatel, Switzerland.

Murphy, N. A., & Hall, J. A. (2011). Intelligence and interpersonal sensitivity: A meta-analysis. *Intelligence*, *39*, 54–63.

*Murry, M. (2014). Unpublished data, Northeastern University.

*Nauts, S., Vonk, R., & Wigboldus, D. H. J. (2010). Unpublished data, Radboud University.

Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review*, *9*, 184–211.

*North, M. S., Todorov, A., & Osherson, D. N. (2012). Accuracy of inferring self- and other-preferences from spontaneous facial expressions. *Journal of Nonverbal Behavior, 36*, 227-233.

*Nowicki, S., Jr., & Duke, M. P. (1992). The association of children's nonverbal decoding abilities with their popularity, locus of control, and academic achievement. *Journal of Genetic Psychology, 15*, 385-393. (counted with Carton & Nowicki, 1993)

Nowicki, S., & Duke, M. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy scale. *Journal of Nonverbal Behavior, 18*, 9-34.

*Nowicki, S., Jr., Glanville, D., & Demertzis, A. (1998). A test of the ability to recognize emotion in the facial expressions of African American Adults. *Journal of Black Psychology, 24*, 335-350.

*Phillips, L. H., MacLean, R. D. J., & Allen, R. (2002). Age and the understanding of emotions: Neuropsychological and sociocognitive perspectives. *Journal of Gerontology: Psychological Sciences, 57B*, P526-P530.

*Pitterman, H., & Nowicki, S., Jr. (2004). A test of the ability to identify emotion in human standing and sitting postures: The Diagnostic Analysis of Nonverbal Accuracy-2 Posture Test (DANVA2-POS). *Genetic, Social, and General Psychology Monographs, 130*, 146-162.

*Rago, S. (2010). Unpublished data, Northeastern University.

*Realo, A., Allik, J., Nõlvak, A., Valk, R., Ruus, T., Schmidt, M., & Eilola, T. (2003). Mind-reading ability: Beliefs and performance. *Journal of Research in Personality, 37*, 420-445.

*Riebe, C. M. (2005). *Emotional intelligence as a predictor of victimization among adolescent males*. Unpublished dissertation, St. John's University.

*Roberts, R. D., Schulze, R., O'Brien, K., MacCann, C., Reid, J., & Maul, A. (2006). Exploring the validity of the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) with established emotions measures. *Emotion, 6*, 663-669.

*Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore: Johns Hopkins University Press. (some results counted with Zuckerman et al., 1975)

Ruben, M. A. (2016). Interpersonal accuracy in the clinical setting. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 287-308). Cambridge, UK: Cambridge University Press.

*Ruben, M. A., Hall, J. A., Curtin, E. M., & Blanch-Hartigan, D. (2015). Discussion

increases efficacy when training accurate perception of patients' affect. *Journal of Applied Social Psychology, 45,* 355-362.

*Ruben, M. A., & Hall, J. A. (2013). "I know your pain": Proximal and distal predictors of pain detection accuracy. *Personality and Social Psychology Bulletin, 39,* 1346-1358.

*Ruben, M. A., Hill, K. M., & Hall, J. A. (2014). How women's sexual orientation guides accuracy of interpersonal judgements of other women. *Cognition & Emotion, 28,* 1512-1521.

*Rule, N. O. (2009). Unpublished data, University of Toronto.

*Rule, N. O. (2010). Unpublished data, University of Toronto.

Rule, N. O., Ambady, N., Adams, R. B., Jr., & Macrae, C. N. (2008). Accuracy and awareness in the perception and categorization of male sexual orientation. *Journal of Personality and Social Psychology, 95*, 1019-1028.

Sayla, G. N., Vella, L., Armstrong, C. C., Penn, D. L., & Twamley, E. W. (2013). Deficits in domains of social cognition in schizophrenia: A meta-analysis of the empirical evidence. *Schizophrenia Bulletin, 39*, 979-992.

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology, 8*, 467-487.

Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion, 1*, 331-346.

*Scherer, K. R., & Scherer, U. (2011). Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the Emotion Recognition Index. *Journal of Nonverbal Behavior, 35*, 305-326. (counted with Bänziger et al., 2009)

*Schlegel, K., & Scherer, K. R. (2015). Introducing a short version of the Geneva

Emotion Recognition Test (GERT-S): Psychometric properties and construct validation. *Behavior Research Methods,* 1-10.

*Schlegel, K., Fontaine, J., & Scherer, K. R. (2014). Unpublished data, University of Geneva.

Schlegel, K., Grandjean, D., & Scherer, K. R. (2012). Emotion recognition: Unidimensional Ability or a set of modality- and emotion-specific skills? *Personality and Individual Differences, 53*, 16-21.

*Schlegel, K., Grandjean, D., & Scherer, K. R. (2013). Constructs of social and emotional effectiveness: Different labels, same content? *Journal of Research in Personality, 47*, 249–253.

Schlegel, K., Grandjean, D., & Scherer, K. R. (2014). Introducing the Geneva Emotion Recognition Test: An example of Rasch-based test development. *Psychological Assessment, 26*, 666-672.

*Schlegel, K., & Jaksic, C. (2014). Unpublished data, University of Geneva.

*Schlegel, K., & Mortillaro, M. (2014). Unpublished data, University of Geneva.

*Schmid, P. C. (2011a). Unpublished data, University of Neuchâtel.

*Schmid, P. C. (2011b). Unpublished data, University of Neuchâtel.

*Schmid, P. C., Schmid Mast, M., Bombari, D., & Mast, F. W. (2011). Gender effects in information processing on a nonverbal decoding task. *Sex Roles, 65*, 102-107.

*Schmid Mast, M., Bangerter, A., Bulliard, C., & Aerni, G. (2011). How accurate are recruiters' first impressions of applicants in employment interviews? *International Journal of Selection and Assessment, 19*, 198-208.

*Schmidt, E. K., Boone, R. T., Isaacowitz, D. M., & Cunningham, J. G. (2011). Unpublished data, Brandeis University.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350-353.

*Schwartz, R. (2014). Unpublished data, McGill University.

*Sternberg, R. J., & Smith, C. (1985). Social intelligence and decoding skills in nonverbal communication. *Social Cognition, 3*, 168-192.

*Stricker, L. J., & Rock, D. A. (1990). Interpersonal competence, social intelligence, and general ability. *Personality and Individual Differences, 8*, 833-839.

Thompson, A. E., & Voyer, D. (2014). Sex differences in the ability to recognize non-verbal displays of emotion: A meta-analysis. *Cognition and Emotion, 28*, 1164-1195.

Thorndike, E.L. (1920). Intelligence and its use. *Harper's Magazine, 140*, 227-235.

*Toomey, R., Seidman, L. J., Lyons, M. J., Faraone, S. V., & Tsuang, M. T. (1999). Poor perception of nonverbal social-emotional cues in relatives of schizophrenic patients. *Schizophrenia Research, 40*, 121-130.

Tskhay, K. O., & Rule, N. O. (2013). Accuracy in categorizing perceptually ambiguous groups: A review and meta-analysis. *Personality and Social Psychology Review, 17*, 72-86.

Uljarevic, M., & Hamilton, A. (2013). Recognition of emotions in autism: A formal meta-analysis. *Journal of Autism and Developmental Disorders, 43*, 1517-1526.

Vogt, D. S., & Colvin, C. R. (2003). Interpersonal orientation and the accuracy of personality judgements. *Journal of Personality, 71*, 267-295.

Vrij, A., & Semin, G. R. (1996). Lie experts' beliefs about nonverbal indicators of deception. *Journal of Nonverbal Behavior, 20,* 65-80.

*Warren, G., Schertler, E., & Bull, P. (2009). Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior, 33*, 59-69.

*Wakabayashi, A., & Katsumata, A. (2011). The Motion Picture Mind-Reading Test: Measuring individual differences of social cognitive ability in a young adult population in Japan. *Journal of Individual Differences, 32*, 55-64.

*Wickline, V. B., Bailey, W., & Nowicki, S. (2009). Cultural in-group advantage: Emotion recognition in African American and European American faces and voices. *Journal of Genetic Psychology, 170*, 5-29.

*Yeagley, E., Morling, B., & Nelson, M. (2007). Nonverbal zero-acquaintance accuracy of self-esteem, social dominance orientation, and satisfaction with life. *Journal of Research in Personality, 41*, 1099-1106.

Zebrowitz, L. A. (2001). Groping for the elephant of interpersonal sensitivity. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 333-350). Mahwah, NJ: Lawrence Erlbaum Associates.

*Zuckerman, M., Hall, J. A., DeFrank, R. S., & Rosenthal, R. (1976). Encoding and decoding of spontaneous and posed facial expressions. *Journal of Personality and Social Psychology, 34*, 966-977.

*Zuckerman, M., Lipets, M. S., Koivumaki, J. H., & Rosenthal, R. (1975). Encoding and decoding nonverbal cues of emotion. *Journal of Personality and Social Psychology, 32*, 1068-1076.

Footnotes

1. We also coded the criterion that the test developer used to determine what the "right" and "wrong" answers were on a given test (e.g., instruction that the target received from test developer or target's self-report) and examined effect sizes as a function of similarity between tests on the criterion. However, this variable was essentially redundant with stimulus creation mode: when stimuli were posed by the target, the "right" answer usually corresponded to the instruction the target had received, such as the instruction to pose a certain emotional expression; when stimuli were spontaneous, the "right" answer usually corresponded to the target's self-report on the variable to be judged. Accuracy criterion is therefore not discussed further.